HOKKAIDO UNIVERSITY

Title	Text-to-image GAN-based Scene Retrieval and Re-ranking Considering Word Importance
Author(s)	Yanagi, Rintaro; Togo, Ren; Ogawa, Takahiro et al.
Citation	IEEE Access, 7(1), 169920-169930 https://doi.org/10.1109/ACCESS.2019.2952676
Issue Date	2019-11-11
Doc URL	https://hdl.handle.net/2115/76281
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	journal article
File Information	08895780.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 18, 2019, accepted October 30, 2019, date of publication November 11, 2019, date of current version December 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2952676

Text-to-Image GAN-Based Scene Retrieval and Re-Ranking Considering Word Importance

RINTARO YANAGI[1], (Student Member, IEEE), REN TOGO[2], (Member, IEEE), TAKAHIRO OGAWA[2], (Senior Member, IEEE), AND MIKI HASEYAMA[2], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Faculty of Information Science and Technology, Division of Media and Network Technologies, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Rintaro Yanagi (yanagi@lmd.ist.hokudai.ac.jp)

This work was supported in part by the MIC/SCOPE under Grant #181601001.

ABSTRACT In this paper, we propose a novel scene retrieval and re-ranking method based on a text-to-image Generative Adversarial Network (GAN). The proposed method generates an image from an input query sentence based on the text-to-image GAN and then retrieves a scene that is the most similar to the generated image. By utilizing the image generated from the input query sentence as a query, we can control semantic information of the query image at the text level. Furthermore, we introduce a novel interactive re-ranking scheme to our retrieval method. Specifically, users can consider the importance of each word within the first input query sentence. Then the proposed method re-generates the query image that reflects the word importance provided by users. By updating the generated query image based on the word importance, it becomes feasible for users to revise retrieval results through this re-ranking process. In experiments, we showed that our retrieval method including the re-ranking scheme outperforms recently proposed retrieval methods.

INDEX TERMS Text-to-image generative adversarial network, multimedia information retrieval, scene retrieval, re-ranking.

I. INTRODUCTION

With the recent exponential growth of Web services such as YouTube[1] and Netflix,[2] the amount of video data has greatly increased. About 72 hours of videos are now uploaded to YouTube every minute [1]. It has become difficult for users to find desired scenes from such a large database. When a user has only a slight recollection of which video contains a desired scene, it takes a long time and much effort for the user to find the desired scene. Therefore, realization of accurate and robust scene retrieval has become an urgent task in the big data era [2]–[6].

Scene retrieval is often achieved by frame-based methods that treat a target database of videos as a set of frames and evaluate each frame [2]–[6]. The frame-based methods are naive but are one of the strong solutions for a scene retrieval task. Frame-based scene retrieval methods can be broadly classified into the following two groups by their input queries, query-by-sentence [2]–[4] and query-by-content [5], [6]. Query-by-sentence methods retrieve an objective scene by comparing an input query sentence and text annotations associated with each frame in a target database. Query-by-sentence methods allow users to retrieve an objective scene easily by utilizing a sentence as an input query. On the other hand, query-by-content methods retrieve an objective scene by comparing visual features respectively extracted from an input content (*e.g.*, an image) and each frame. Accurate text annotations do not have to be prepared for these methods since visual features are used for the calculation of retrieval ranking. In addition, content information can be directly used in query-by-content methods. If high-level semantic features can be extracted from a query image, query-by-content methods can retrieve an objective scene with high performance. Query-by-sentence methods and query-by-content methods therefore both have advantages.

Although high-performance retrieval is realized by each retrieval method, query-by-sentence and query-by-content approaches cannot retrieve a desired scene from

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

[1]https://www.youtube.com
[2]https://www.netflix.com

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

an inappropriate query. The quality of a query has a significant influence on the scene retrieval performance [7], [8]. In this regard, there remain several problems in query-by-sentence and query-by-content retrieval methods. First, query-by-sentence methods require text annotations on target video data to compare an input query sentence and candidate scenes. In other words, retrieval performance greatly depends on the quality of the text annotations in the target database. If a sentence similar to the query is not assigned to the desired scene in the target database, users cannot obtain the desired scene by a query-by-sentence method [5]–[7], [9], [10]. Since providing accurate annotation for each scene is a labor-intensive process, it is difficult for users to give specific text annotations with the explosive increase in the number of videos [11], [12]. Second, although query-by-content methods can retrieve scenes without text annotations on the database, it is often impossible to prepare a suitable query content for retrieval. Hence, the situation in which we have an appropriate query content is not suitable for real-world scene retrieval situations. In other words, situations in which query-by-content methods can be used are limited.

Generally, it is difficult for a user to correct retrieval results to the desired scene when the retrieval result obtained by the first input query is not the user's desired scene. For example, if a user wants to retrieve a scene that contains a dog and a person, the user can input the query "A man and a dog". If the retrieval result is not a desired scene (*e.g.*, it contains only a dog), the user cannot revise the retrieval result to the desired scene. To solve this problem, a new retrieval method that can re-rank retrieval results from user feedback is required. Robust re-ranking of retrieval results from user feedback is a challenging task; however, various re-ranking methods that can improve retrieval performance have been proposed [13]–[16].

In this paper, we propose a new retrieval method that can solve the above mentioned problems. The proposed retrieval method is built on a text-to-image generative adversarial network (GAN) [17]. Text-to-image GANs [17]–[24] are deep learning techniques that can generate sample images from an input query sentence via several trained neural networks adversarially. Text-to-image GANs can generate images that represent an input query sentence, and the core idea of our method is this multimodal translation. In the proposed method, a query sentence is firstly inputted to the text-to-image GAN. Next, a synthetic image corresponding to the input query sentence is generated through the text-to-image GAN. The generated image is used as a query image for the scene retrieval task. By utilizing the image generated from an input query sentence, we can easily control semantic information of the query image at the text level, and we can retrieve an objective scene without relying on text annotation. Furthermore, we introduce a novel interactive re-ranking scheme to the above-mentioned retrieval method based on the structure of the GAN. We give a weight for each word feature of an input query sentence and update the generated image based on these weighted word features. By revising the generated image, a user can search for an objective scene while adjusting the word importance of the first input query sentence. This is the most novel scheme in the proposed method. Based on these schemes, high performance and more robust scene retrieval is realized.

The major contributions of our study are summarized as follows.

Retrieval without relying on text annotation
Retrieval that does not require text annotations on a target database is realized by introducing a text-to-image GAN into the scene retrieval task. From this mechanism, our method allows users to input diverse queries as a form of text.

Easy manipulation of a query at the semantic level
We can control semantic information of the query image from the text level, and then our method can compare an input query sentence and each frame in a target database on visual feature space.

Re-ranking considering word importance
When a user obtains an undesired scene as the first retrieval result, the user can revise the first retrieval result by introducing a scheme that can update the generated image based on the word importance of the input query sentence.

II. RELATED WORKS

In this section, we describe studies related to our method. In Subsection II-A, we describe related studies on multimedia information retrieval and explain the differences between these methods and our method. In Subsection II-B, we describe studies on text-to-image GANs and their characteristics.

A. MULTIMEDIA INFORMATION RETRIEVAL

The scene retrieval task from an input query sentence can be considered as one of retrieval tasks that match an input query sentence and contents in a target database. The mainstream of this multimodal retrieval task can be roughly divided into two types: methods that utilize text annotations and methods that embed visual and sentence features into common semantic spaces.

First, we describe retrieval methods that utilize text annotations provided from candidate contents [16], [25]–[27]. These methods retrieve objective contents by comparing an input query sentence and text annotations attached to the candidate contents. Various deep learning-based methods to automatically provide text annotations for candidate contents have been proposed in recent years. Karpathy and Li [25] proposed a method in which text annotations are given to corresponding subregions of candidate images by utilizing a Convolutional Neural Network (CNN) and bidirectional Recurrent Neural Network (bRNN) [28]. Since this method gives annotations not to candidate images but to subregions of candidate images, it is possible to retrieve objective images even if these images contain various objects. Vinyals *et al.* [26] estimated a

long sentence using Long Short-Term Memory (LSTM) [29]. Their method can conduct retrieval that considers the relationships between objects in a candidate image based on the estimated long sentence. Unlike the above-described methods, our method does not utilize text annotations.

Next, we describe retrieval methods that embed visual and sentence features into common semantic spaces [7], [9], [10], [30]. These methods retrieve contents by comparing embedded visual and textual features, and then they can compare an input query sentence and candidate contents at the semantic level. Kiros et al. utilized a CNN and LSTM to embed visual and textual features into a common semantic space and retrieve relevant images [9]. Retrieval is realized by this method without using text annotations. Vendrov et al. proposed a method in which visual and textual features are embedded into a common semantic space with consideration of word relationships of an input query sentence [7]. This method realizes retrieval that can consider word relationships of the input query sentence even though the method does not utilize text annotations. These methods have robustness for inputting query sentences since they do not utilize specific text annotations associated with candidate contents. Our retrieval method can be regarded as an extension of these methods.

Various methods that can re-rank retrieval results have been proposed in recent years [13]–[16]. These methods boost retrieval performance by utilizing user intention based on their behavior. For example, Yang et al. proposed a method for re-ranking retrieval results according to the number of clicks on the Web [14]. Lu et al. proposed a method for re-ranking retrieval results based on the user's information and image view times [16]. Since these methods improve retrieval performance through searching behavior of the user, they can conduct retrieval that reflects the user's intention.

Following these re-ranking methods, by focusing on the text-to-image GAN, we construct a novel re-ranking scheme that can update retrieval results based on word importance provided by the user.

B. TEXT-TO-IMAGE GENERATIVE ADVERSARIAL NETWORK AND ITS APPLICATION

Text-to-image GANs have been widely studied by many researchers in recent years. A text-to-image GAN is one of the deep-learning methods, and it can generate an image that reflects information of an input query sentence by training its generator and discriminator alternately. A generator generates images from an input query sentence, and a discriminator discriminates whether input images are those from real data or those generated by the generator. A text-to-image GAN can generate images that reflect the information of the input query sentence by utilizing these different networks. Read et al. proposed the first model in which a GAN was applied to a text-to-image task [19]. Although the generated images are not visually pleasant and the size of the generated images is limited to 64×64 pixels, they showed a new way to use a GAN. Following this model, Zhang et al. proposed a model that can generate more visually pleasant images [24]. This model enables generation of not only visually pleasant images but also more high resolution (256×256 pixels) images. AttnGAN [17] and HDGAN [22] are the latest models in this field. AttnGAN utilizes word features in addition to sentence features that are normally utilized by other text-to-image GANs. This structure strongly matches the aim of our study, that is, re-ranking retrieval results based on the word importance of the input query sentence provided by the user. We realize a retrieval model that can reflect user intention by focusing on this structure.

Here, there are various methods applying GAN for information retrieval tasks that retrieve target documents from a query sentence [31], [32]. They utilize adversarial relationship between a generator and a discriminator to calculate relevance of candidate documents and a query sentence. Different from these papers, we utilize text-to-image GAN as a generator that generates a query image from a query sentence and we retrieve a target scene related to the generated query image.

III. SCENE RETRIEVAL AND RE-RANKING METHOD

In this section, we present the details of the proposed method. Figure 1 shows an overview of our retrieval approach, and, Fig. 2 shows an overview of our re-ranking scheme. Our method consists of three steps (A-C shown in Figs. 1 and 2): query image generation, scene retrieval from a generated query image and re-ranking by query image re-generation. In the first step, we construct a text-to-image translation network and generate a query image from a user input query sentence. In the second step, our method retrieves scenes that are similar to the generated query image. These first and second steps correspond to the left (A) and right (B) parts in Fig. 1, respectively. If the user is not satisfied with the retrieval result after the second step, the user conducts the third step. In the third step, the user provides the word importance to the input query sentence considering the excess or shortage of the first retrieval results. Based on the word importance, our method re-generates a query image and retrieves a scene similar to the re-generated image. This step corresponds to C in Fig. 2.

In Subsection III-A, we describe how to generate an image that reflects the information of the input query sentence. In Subsection III-B, we describe the retrieval of the objective scene by using the generated image. The re-ranking scheme using the query image re-generation is presented in Subsection III-C.

A. QUERY IMAGE GENERATION

In the first step, the proposed method generates a query image from a user input query sentence. Based on the query-by-content method, the proposed method retrieves a scene similar to the query image generated from the input query sentence. To generate the query image, we construct a text-to-image translation network that has one generation network G and three neural networks N_i ($i = 1, 2, 3$) based on AttnGAN [17]. N_i is a neural network that calculates hidden

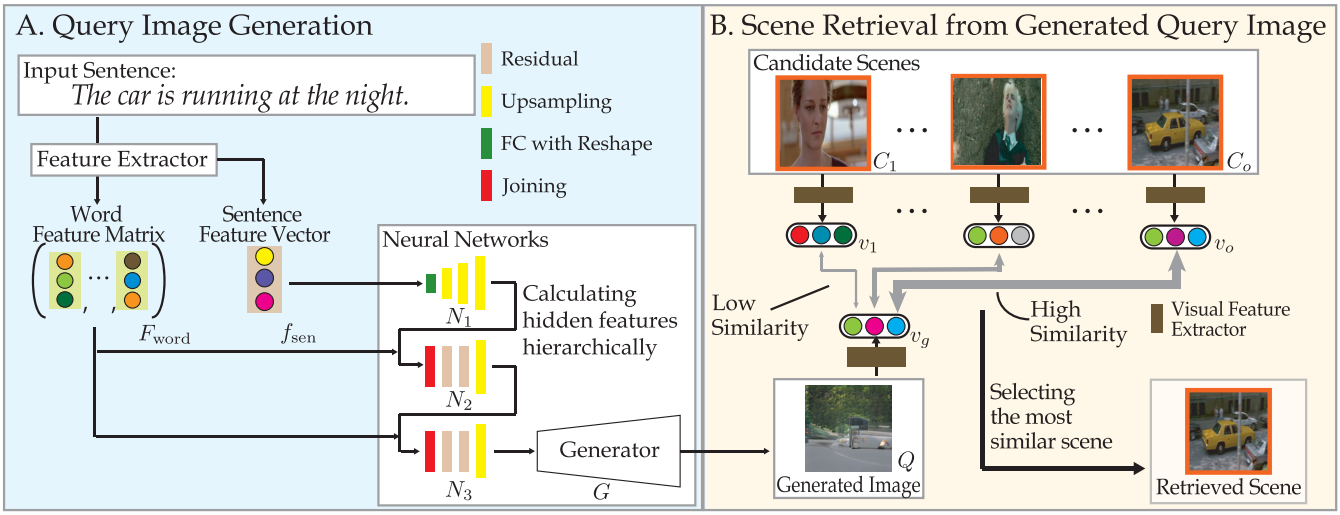


FIGURE 1. Overview of the proposed retrieval approach. The two stages (A and B) are described in detail in III-A and III-B, respectively.

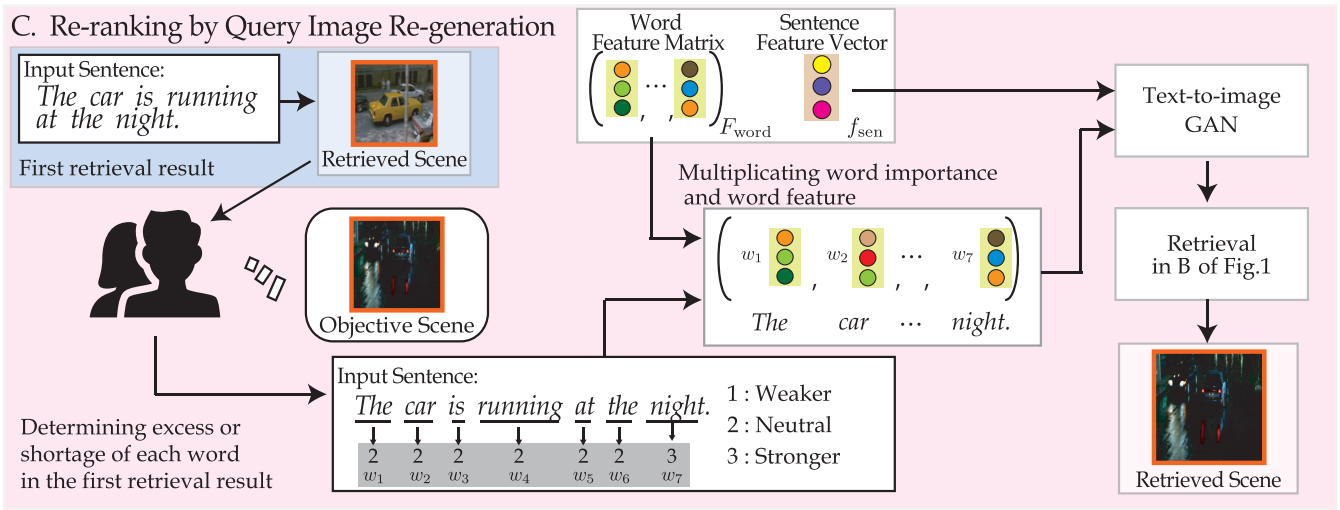


FIGURE 2. Overview of the proposed re-ranking scheme. From the first retrieval result obtained by the retrieval approach shown in Fig. 1, a user gives the importance for each word to obtain its re-ranked result. The details are given in III-C.

vector $\boldsymbol{v}_i$ containing information of an input query sentence, and G is a generation network that generates a 256×256 pixel image Q from hidden vector $\boldsymbol{v}_3$.

First, we define a sentence feature extracted from an input query sentence as $\boldsymbol{f}_{sen} \in \mathbb{R}^{D_{sen}}$ and a word feature extracted from the nth word of an input query sentence as $\boldsymbol{f}^n_{word} \in \mathbb{R}^{D_{word}}$ ($n = 1, 2, \ldots, N$; N being the number of words included in the sentence). $\boldsymbol{f}_{sen}$ and $\boldsymbol{f}^n_{word}$ are calculated from a bi-directional Long Short-Term Memory (LSTM) [28] described in [17]. Here, D_{sen} and D_{word} represent the dimensions of the extracted sentence and word features, respectively. The feature $\boldsymbol{f}_{sen}$ contains information on the relationships between the words of the input query sentence, and the feature $\boldsymbol{f}^n_{word}$ contains information on each word. Then we construct a word feature matrix as $\boldsymbol{F}_{word} = (\boldsymbol{f}^1_{word}, \boldsymbol{f}^2_{word}, \cdots, \boldsymbol{f}^N_{word}) \in \mathbb{R}^{D_{word} \times N}$. From the sentence feature $\boldsymbol{f}_{sen}$ and the Gaussian noise $\boldsymbol{z}$, we obtain the hidden vector $\boldsymbol{v}_1$ as follows:

$$\boldsymbol{v}_1 = N_1(\boldsymbol{z}, N^{ca}(\boldsymbol{f}_{sen})), \tag{1}$$

where N^{ca} is a neural network that stabilizes the training [24] and translates $\boldsymbol{f}_{sen}$ to continuous features. Next, we calculate the hidden vectors $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ as follows:

$$\boldsymbol{v}_2 = N_2(\boldsymbol{v}_1, N_2^{attn}(\boldsymbol{F}_{word}, \boldsymbol{v}_1)), \tag{2}$$

$$\boldsymbol{v}_3 = N_3(\boldsymbol{v}_2, N_3^{attn}(\boldsymbol{F}_{word}, \boldsymbol{v}_2)), \tag{3}$$

where N_i^{attn} ($i = 2, 3$) integrates the word feature matrix $\boldsymbol{F}_{word}$ into the previous hidden vector $\boldsymbol{v}_i$ ($i = 1, 2$). We obtain the hidden vector $\boldsymbol{v}_2$ from vector $\boldsymbol{v}_1$ and the word feature matrix $\boldsymbol{F}_{word}$. Hidden vector $\boldsymbol{v}_2$ contains information on the sentence features $\boldsymbol{f}_{sen}$ and contains some information on the

word feature matrix $\boldsymbol{F}_{word}$. Similarly, we obtain hidden vector $\boldsymbol{v}_3$ from vector $\boldsymbol{v}_2$ and the word feature matrix $\boldsymbol{F}_{word}$. Hidden vector $\boldsymbol{v}_3$ contains information on both the sentence and word feature matrix. Our text-to-image translation network can generate images that focus on each word through this function. Finally, we generate the image $Q = G(\boldsymbol{v}_3)$ from hidden vector $\boldsymbol{v}_3$. In the proposed method, we utilize this generated image Q to retrieve relevant scenes in the following step.

Here, we show the training strategy of our text-to-image translation network to generate a query image that contains information of an input query sentence. Our text-to-image translation network is trained to minimize the loss function defined as follows:

$$L = L_G + \lambda L_{\mathrm{DAMSM}}, \tag{4}$$

where λ balances L_G and L_{DAMSM}. Specifically, L_G is a loss function that approximates conditional and unconditional distributions. L_{DAMSM} measures the image text similarity at the word level by embedding subregions of the image and words of the sentence to a common semantic space. L_{DAMSM} is described in detail in [17]. Also, discriminator D is trained to classify whether the input image is real or fake.

B. SCENE RETRIEVAL FROM A GENERATED QUERY IMAGE

In the second step, we retrieve a scene that is the most similar to the image Q generated in the first step. By utilizing the generated image that represents the information of the input query sentence, the proposed method can retrieve scenes in visual feature space while we utilize a sentence as the input. We adopt a simple query-by-content method for retrieving an objective scene. First, we calculate visual features $\boldsymbol{v}_q$ and $\boldsymbol{v}_p$ from Q and candidate frames C_p ($p = 1, 2, \ldots, P$; P being the number of frames included in all candidate scenes). Here, we utilize outputs of the third pooling layer of Inception-v3 [33] pre-trained on ImageNet [34] as the visual features because the loss function L_{DAMSM} utilizes Inception-v3 as the image feature extractor for calculating the image-text matching loss. Then we calculate similarities s_p between $\boldsymbol{v}_q$ and $\boldsymbol{v}_p$ as follows:

$$s_p = \frac{\boldsymbol{v}_q \cdot \boldsymbol{v}_p}{|\boldsymbol{v}_q||\boldsymbol{v}_p|} \quad (p = 1, 2, \cdots, P). \tag{5}$$

Namely, s_p represents the similarity between the generated image Q and candidate frames C_p, and high s_p indicates that Q and C_p contain similar information. Finally, we obtain the scene containing the frame that has higher similarity than that of other frames as the retrieval result.

C. RE-RANKING BY QUERY IMAGE RE-GENERATION

This step becomes necessary when the estimated scene is not a user's desired scene. As shown in Fig. 2, we re-generate the query image $\hat{Q}$ based on each word importance provided by the user and revise the retrieval result by utilizing the re-generated image. First, the user considers the excess or shortage of each word in the first retrieval result and sets the word importance w_n ($n = 1, 2, \cdots, N$) for each word as shown in Fig. 2. Then the proposed method calculates $\boldsymbol{F}^{re}_{\mathrm{word}}$ as follows:

$$\boldsymbol{F}^{re}_{\mathrm{word}} = (w_1 \boldsymbol{f}^1_{\mathrm{word}}, w_2 \boldsymbol{f}^2_{\mathrm{word}}, \cdots, w_N \boldsymbol{f}^N_{\mathrm{word}}). \tag{6}$$

By strengthening or weakening a word feature that represents information of a specific word than other word features, the user can change the degree of expression of each word in $\boldsymbol{F}_{\mathrm{word}}$. Namely, the user can adjust the proportion of word information represented in the query. Finally, the proposed method re-generates an image $\hat{Q}$ utilizing $\boldsymbol{F}^{re}_{\mathrm{word}}$ based on the method described in Subsection III-A and obtains the re-ranking result by performing the retrieval shown in Subsection III-B again.

We show an example of the re-ranking situation. If a user inputs a query "A man and a dog", a query image is generated under a setting of the word importance $\{w_1, w_2, w_3, w_4, w_5\} = \{1.0, 1.0, 1.0, 1.0, 1.0\}$ and the proposed method retrieves a scene based on the generated image. Here, w_1, w_2, w_3, w_4, w_5 are word importance of "A", "man", "and", "a", "dog", respectively. However, if the user obtains a scene that contains only a dog, the user wants to emphasize the word "man". Therefore, the user sets the word importance as $\{w_1, w_2, w_3, w_4, w_5\} = \{1.0, 2.0, 1.0, 1.0, 1.0\}$, and the proposed method performs the re-ranking.

This re-ranking scheme enables a user to retrieve a user's desired scene by adjusting the importance of each word. Specifically, a user can modify a generated image to an image that can retrieve a user's desired scene with watching the retrieval results, and then the user can obtain the retrieval result matched to the user's desired scene. This means that our method can reflect the user's intention in the retrieval result effectively through the re-ranking scheme.

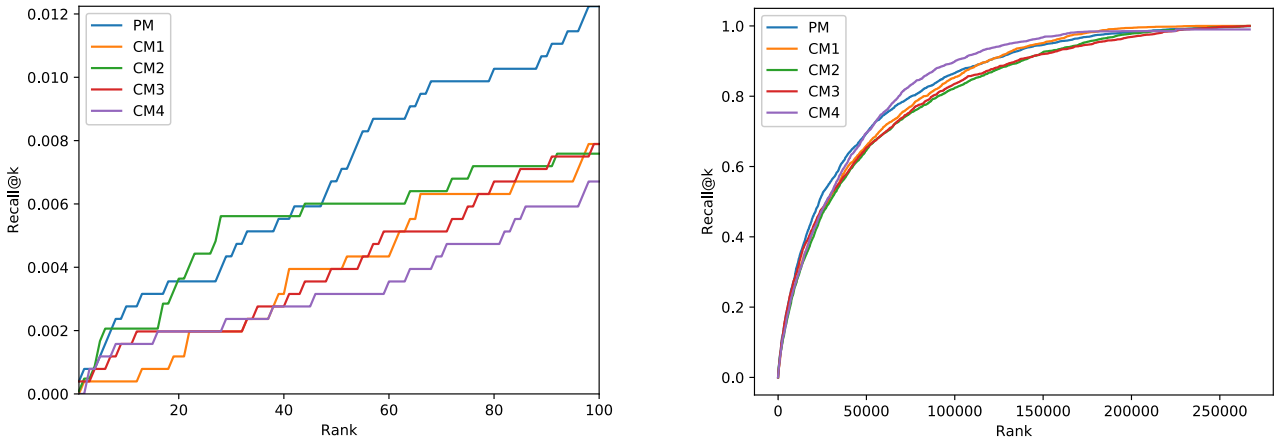
IV. EXPERIMENTAL RESULTS

In this section, we describe the settings and results if experiments using some popular large-scale video datasets. We conducted three experiments to evaluate our retrieval performance and re-ranking performance. In the first experiment, we compared the scene retrieval performance of our method with the performance of some state-of-the-art methods. In the second experiment, we evaluated the re-ranking performance of our method by a subjective experiment. Finally, we confirmed the re-ranking performance.

Details of the dataset used in this experiment are shown in Subsection IV-A. In Subsection IV-B, we show our retrieval performance. Qualitative and quantitative results of the re-ranking scheme are presented in Subsection IV-C.

A. DATASETS USED IN EXPERIMENTS

Scene retrieval performance heavily relies on the contents of the dataset and their annotations. Hence, robustness of a scene retrieval method should be evaluated with some different types of video datasets. Since a scene retrieval task targets various scenes, the text-to-image GAN should be trained



(a) Recall@k (k= 1,2,⋯,100)

(b) Recall@k (k= 1,2,⋯,P;P= 267,320)

FIGURE 3. Recall@k obtained from the movies "As Good As it Gets", "Bad Santa", "Halloween", "Harry Potter and the Prisoner of Azkaban" and "Rendezvous mit Joe Black" from the MP-II MD dataset.

on the dataset that can represent such a situation. In this study, we utilized a large-scale Microsoft Common Objects in Context (MSCOCO) dataset [35] for training our text-to-image translation network. The MSCOCO dataset consists of images and text annotations that express information on these images. This dataset is often used for object recognition, segmentation and multimodal translation tasks. In the proposed method, we trained our text-to-image translation network using 82,783 training images and text annotations attached to each training image.

We assume that our scene retrieval method can be used for a dataset without any annotations such as new movies, home videos and lifelog videos. Therefore, we utilized the following two datasets to evaluate the retrieval performance in the above-mentioned situation.

MP-II Movie Description (MP-II MD) dataset [36]
The MP-II MD dataset contains 68,000 scenes from 94 HD movies, and each scene is associated with one description. In this experiment, we utilized their descriptions as input query sentences for generating query images and we defined scenes corresponding to the descriptions as ground truth. We randomly selected 5 movies for evaluation: "As Good As it Gets", "Bad Santa", "Halloween", "Harry Potter and the Prisoner of Azkaban" and "Rendezvous mit Joe Black" (2, 532 scenes and 267, 320 frames).

Charades-Ego dataset [37]
The Charades-Ego dataset consists of 7,860 scenes (580, 570 frames) of daily indoor activities and an annotation for each scene collected through Amazon Mechanical Turk.[3] Each scene is recorded from both the third-person perspective and the first-person perspective. In this experiment, we utilized scenes recorded from the first-person perspective and text annotations as input query sentences for generating query images and we defined the scene corresponding to the annotation as the ground truth.

[3]https://www.mturk.com/

TABLE 1. MAP obtained from the movies "As Good As it Gets", "Bad Santa", "Halloween", "Harry Potter and the Prisoner of Azkaban" and "Rendezvous mit Joe Black" from the MP-II MD dataset.

	PM	CM1	CM2
MAP	**0.0164**	0.00545	0.0124
	CM3	CM4	
MAP	0.00902	0.00646	

TABLE 2. MAP obtained from the Charades-Ego dataset.

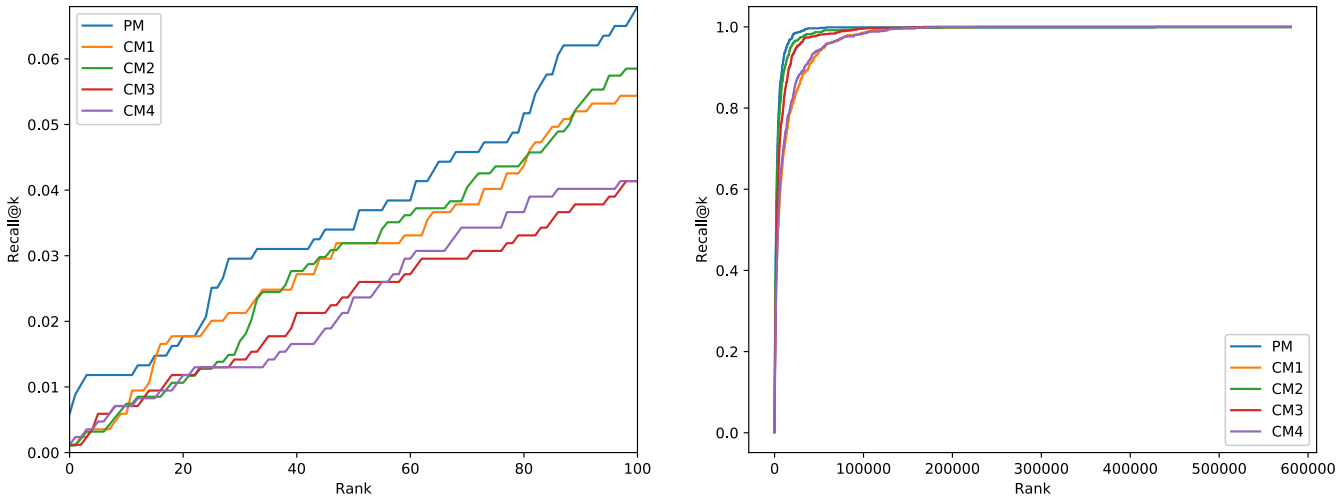
	PM	CM1	CM2
MAP	**0.0815**	0.0662	0.0623
	CM3	CM4	
MAP	0.0527	0.0564	

B. RETRIEVAL PERFORMANCE EVALUATION

We show the retrieval performance of our method. We utilized all text annotations attached to each dataset as inputs. We defined frames included in the target scene as ground truth. And then we calculated the Recall@k for evaluating retrieval performance following previous multimodal retrieval methods [7], [9], [10], [30] and calculated Mean Average Precision (MAP). Recall@k and MAP are defined as follows:

$$\text{Recall@}k = \frac{t_k}{M} \quad (k = 1, 2, \ldots, P), \tag{7}$$

$$\text{MAP} = \frac{1}{M}\sum_{m=1}^{M} \text{AP}_m, \tag{8}$$



(a) Recall@k (k= 1,2,⋯,100)

(b) Recall@k (k= 1,2,⋯,P;P= 267,320)

FIGURE 4. Recall@k obtained from the Charades-Ego dataset.

TABLE 3. Results of the subjective experiment for evaluating the re-ranking performance quantitatively. "Retrieval result" represents the average score given by each subject for evaluation of the first retrieval results and "Re-ranked result" represents the average score given by each subject for evaluation of the re-ranked retrieval results. "1" expresses "Not Related" and "5" expresses "Related".

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10
Retrieval result	2.20	2.95	2.00	2.83	2.42	2.32	3.30	2.00	2.85	3.40
Re-ranked result	**3.10**	**3.53**	**2.20**	**3.63**	**3.41**	**3.25**	**3.78**	**3.25**	**3.43**	**4.10**

	Subject 11	Subject 12	Subject 13	Subject 14	Subject 15	Average
Retrieval result	3.12	2.73	2.54	3.11	2.10	2.66
Re-ranked result	**3.79**	**3.17**	**3.32**	**3.62**	**3.12**	**3.38**

$$\mathrm{AP}_m = \frac{1}{C_P}\sum_{p=1}^{P}\alpha_p\frac{C_p}{p}, \tag{9}$$

where t_k is the number of the correctly retrieved scenes in the top-k retrieval results, and M is the number of candidate scenes. According to their similarity ranks, we sorted all P frames in M candidate scenes. Furthermore, we regarded a target scene as being correctly retrieved when the frames of the target scene were included in the top-k retrieval results. Also, AP_m is average precision of mth query ($m = 1, 2, \ldots, M$). Here, C_p is the number of frames in mth scene within p retrieval results, and α_p is binary indicator whether pth retrieval result is correct or not.

We compared the performance of the proposed method (PM) with the performances of some state-of-the-art methods.

- **Comparative Method 1 (CM 1) [9]**
  A deep learning method that embeds visual and textual features into a common semantic feature space utilizing a CNN and LSTM.
- **Comparative Method 2 (CM 2) [7]**
  A deep learning method that considers word relationships of an input query sentence in addition to CM 1. We evaluated whether the proposed method can consider word relationships of an input query sentence.
- **Comparative Method 3 (CM 3) [10]**
  A deep learning method that embeds textual features into a visual feature space. We evaluated whether the proposed method can extract effective visual features.
- **Comparative Method 4 (CM 4) [30]**
  A deep learning method that utilizes a loss function that reduces the number of false samples between a query and an objective sample in addition to CM 1.

Results of overall Recall@k ($k = 1, 2, \cdots, P$) and detailed Recall@k ($k = 1, 2, \cdots, 100$) are shown in Figs. 3 and 4 and MAP is shown in Tables 1 and 2. Here, we utilized 2, 532 query sentences in MP-II MD dataset and 7, 860 query sentences in Charades-Ego dataset, respectively. From Figs. 3 and 4 and Tables 1 and 2, we can see that the proposed method tends to retrieve objective scenes successfully compared to other comparative methods. Specifically, by comparing with CM 2, we can confirm that the proposed method can retrieve objective scenes considering the word relationships of an input query sentence. Also, by comparing with CM 3, we can confirm that generating an image from textual features is more stable than embedding textual features into a visual feature space. On the other hand, unlike

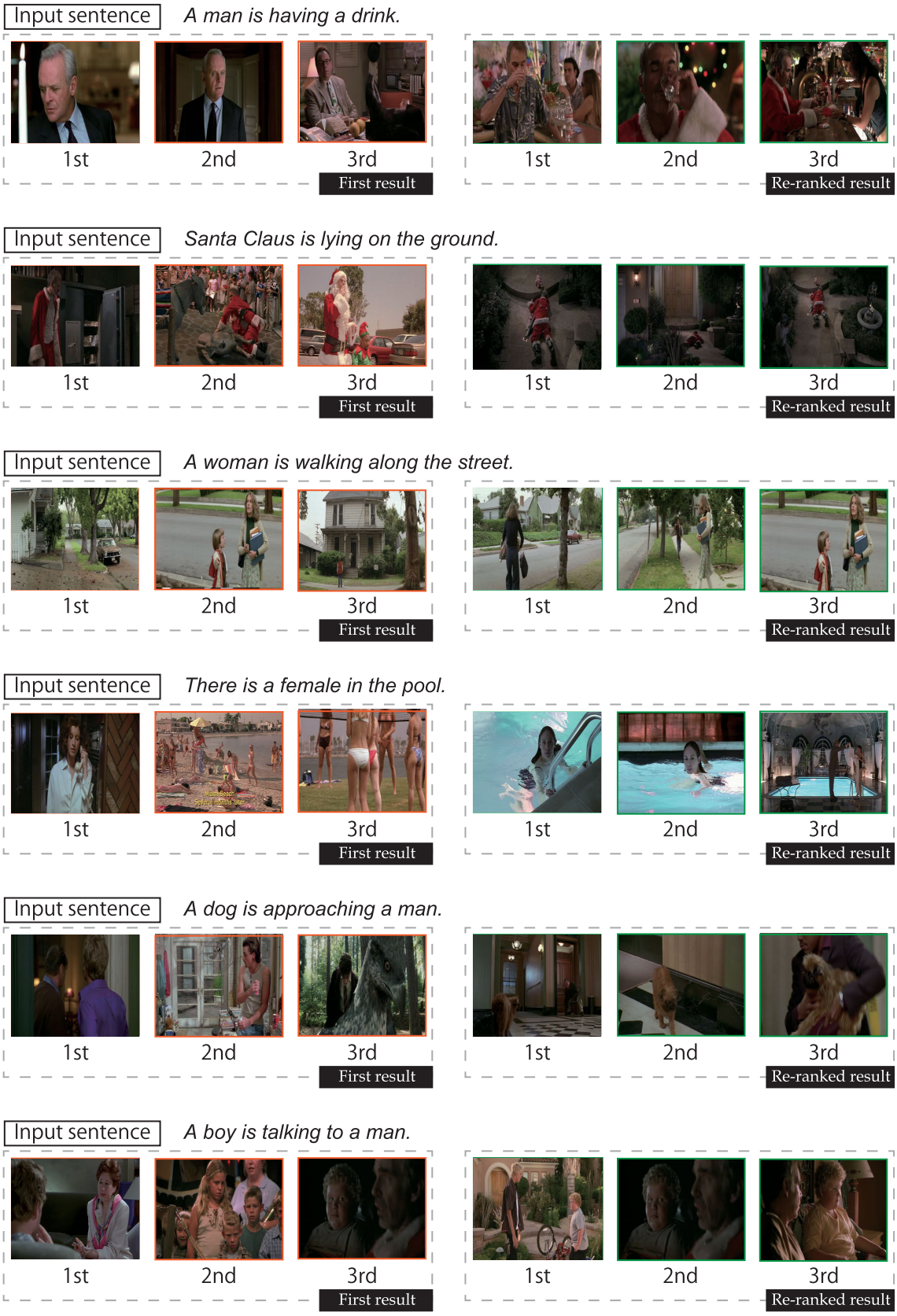
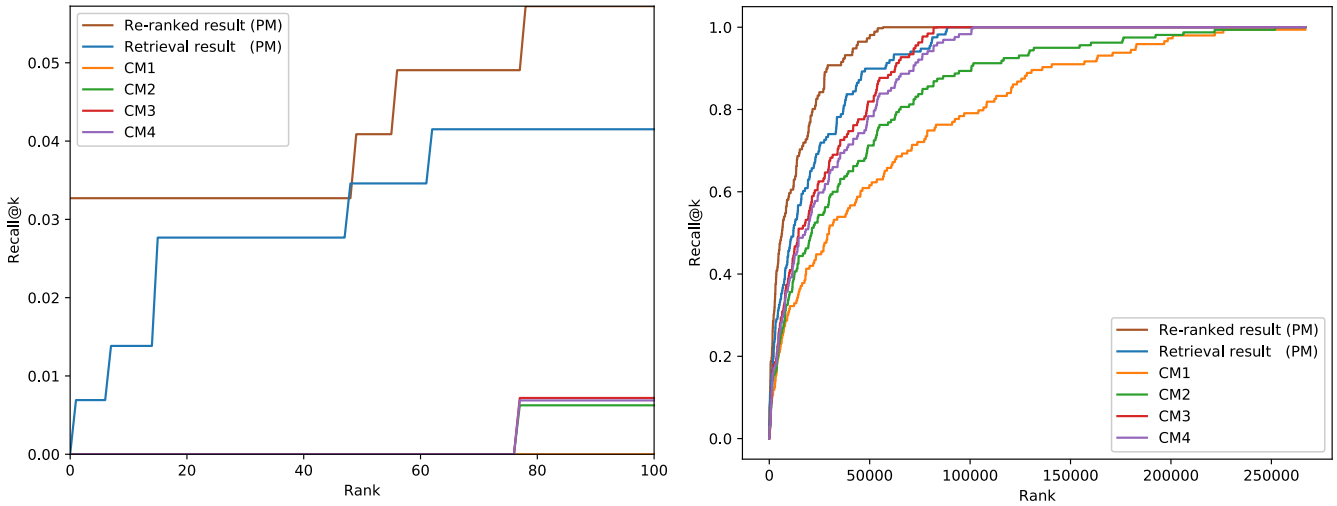


FIGURE 5. Examples of top-3 retrieval and re-ranked results.



(a) Recall@k (k= 1,2,⋯,100)

(b) Recall@k (k= 1,2,⋯,P;P= 267,320)

FIGURE 6. Recall@k obtained in a subjective evaluation.

other comparative methods, the advantage of the proposed method is that the method can revise the retrieval results by considering the word importance. It is expected that the proposed method almost outperforms the other comparative methods by conducting re-ranking described in Sub. III-C.

C. RE-RANKING PERFORMANCE EVALUATION

We conducted a subjective experiment to evaluate the re-ranking performance qualitatively and quantitatively. In this experiment, subjects watched various scenes and considered a sentence to retrieve each scene. We regarded each scene watched by a subject as the ground truth of the sentence that is considered for retrieving each scene. At first, we randomly selected 20 scenes from the MP-II MD dataset. Each subject watched these 20 selected scenes and considered a sentence to retrieve each scene. The subjects were told to avoid utilizing a specific proper noun since it is not the main focus of the proposed method. After that, by utilizing each sentence as the input, we retrieved scenes from the MP-II MD dataset. And then each subject evaluated the matching rates between the retrieval results and each input query sentence and gave the word importance for the re-ranking. The subjects gave the word importance for each word of the input query sentence between one to three levels ("1 Weaker", "2 Neutral", "3 Stronger"). We utilized $w_n = 0.5$ for "1 Weaker", $w_n = 1.0$ for "2 Neutral" and $w_n = 2.0$ for "3 Stronger". Here, w_n is the nth word importance of the input query sentence. Finally, we retrieved scenes from the dataset utilizing each sentence and the word importance, and each subject evaluated the matching rates between the re-ranked results and each input query sentences. In this experiment, 15 subjects (4 females and 11 males, 21-27 years old) evaluated 300 different scenes (20 scenes per subject). For evaluating the matching rate between the retrieval result and the input query sentence, subjects gave an evaluation score of 1 to 5 ("1: Not Related", "2: Not So Related", "3: Neither Agree Nor Disagree", "4: A Little Related" and "5: Related") for each result.

TABLE 4. MAP obtained in a subjective evaluation.

	Re-ranked result (PM)	Retrieval result (PM)	CM1
MAP	**0.178**	0.0959	0.0334
	CM2	CM3	CM4
MAP	0.0412	0.0473	0.0456

The average evaluation scores given by each subject for the first retrieval results and the re-ranked retrieval results are shown in Table 3 and retrieved examples are shown in Fig. 5. In Table 3, "Retrieval result" is the average score given by each subject for evaluation of the first retrieval results and "Re-ranked result" is the average score given for evaluation of the re-ranked retrieval results. It can be seen that the scores for the re-ranked results are better than those for the retrieval results. Thus, it is qualitatively verified that the proposed method can re-rank the first retrieval results.

Next, we calculated Recall@k and MAP to compare retrieval performance of the first retrieval results and the re-ranked results quantitatively. Also, we calculated Recall@k and MAP utilizing other comparative methods described in Subsection IV-B to verify the effectiveness of the re-ranking. In this experiment, to calculate Recall@k and MAP, we used the input query sentences and the word importance prepared for the qualitative evaluation. Here, we utilized 300 query sentences for evaluation. Results of evaluation in the quantitative experiment are shown in Fig. 6 and Table 4. In Fig. 6 and Table 4, "Re-ranked result (PM)" represents the retrieval results utilizing the input query sentences and the word importance. We can see that the retrieval performance of "Re-ranked result (PM)"

outperforms "Retrieval result (PM)". Since "Retrieval result (PM)" is a method that only utilizes the input query sentence, it can be said that the re-ranking is effective for improving the retrieval performance. Also, we can see that the retrieval performance of "Re-ranked result (PM)" greatly outperforms the retrieval performance of other comparative methods.

V. DISCUSSION

Although the proposed method can retrieve an objective scene accurately, there are several limitations to be considered for improving the retrieval performance. For example, we can improve the retrieval performance of the proposed method by considering how to handle a candidate scene in a target database, although a frame-based method is effective for scene retrieval. By considering the action in a candidate scene, it seems that the proposed method effectively utilizes information of a verb in an input query sentence. We will introduce a scheme that can consider the action in the candidate scene into the proposed method in our future work. Furthermore, the scores for both the retrieval result and re-ranked result shown in Table 3 are not so high. In the proposed method, the quality of the generated images directly affects the retrieval performance. This can be solved with the development of a text-to-image GAN. In addition, although there are various methods that can evaluate word importance in text retrieval [38]–[41], we only utilize word importance determined by a user for re-ranking in the proposed method. We will consider the scheme that can automatically determine word importance for re-ranking based on the previous automatic methods such as text statistics.

VI. CONCLUSION

In this paper, we have proposed a novel retrieval and re-ranking method based on a text-to-image GAN. By using the proposed method, a generated image query can be updated and an objective scene can be retrieved by finely adjusting the retrieval result when a user obtains an undesired scene as the first retrieval result. Experimental results showed the effectiveness of the proposed method.

REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[2] M. Ioka and M. Kurokawa, "Estimation of motion vectors and their application to scene retrieval," *Mach. Vis. Appl.*, vol. 7, no. 3, pp. 199–208, Sep. 1994. [Online]. Available: http://link.springer.com/10.1007/BF01211664

[3] Y. Rui, S.-F. Chang, and T. S. Huang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[4] T. Masuda, D. Yamamoto, S. Ohira, and K. Nagao, "Video scene retrieval using online video annotation," in *Proc. Annu. Conf. Jpn. Soc. Artif. Intell.*, 2007, pp. 54–62.

[5] A. Anjulan, N. Canagarajah, M. V. Building, and W. Road, "Video scene retrieval based on local region features," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 3177–3180.

[6] H.-W. Yoo and S.-B. Cho, "Video scene retrieval with interactive genetic algorithm," *Multimedia Tools Appl.*, vol. 34, no. 3, pp. 317–336, 2007.

[7] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–12.

[8] S. Ercoli, M. Bertini, and A. Del Bimbo, "Compact hash codes for efficient visual descriptors retrieval in large scale databases," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2521–2532, Nov. 2017.

[9] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: https://arxiv.org/abs/1411.2539

[10] J. Dong, X. Li, and C. G. M. Snoek, "Word2VisualVec: Image and video to sentence matching by visual feature prediction," 2016, *arXiv:1604.06838*. [Online]. Available: https://arxiv.org/abs/1604.06838

[11] S. Park, H. Park, and C. D. Yoo, "Complex video scene analysis using kernelized-collaborative behavior pattern learning based on hierarchical representative object behaviors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1275–1289, Jun. 2017.

[12] Q. Abbas, M. E. A. Ibrahim, and M. A. Jaffar, "Video scene analysis: An overview and challenges on deep learning algorithms," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20415–20453, 2018.

[13] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," 2017, *arXiv:1708.04169*. [Online]. Available: https://arxiv.org/abs/1708.04169

[14] X. Yang, T. Mei, Y. Zhang, J. Liu, and S. Satoh, "Web image search re-ranking with click-based similarity and typicality," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4617–4630, Oct. 2016.

[15] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 3652–3661.

[16] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.

[17] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[19] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: https://arxiv.org/abs/1605.05396

[20] N. Bodla, G. Hua, and R. Chellappa, "Semi-supervised FusedGAN for conditional image generation," 2018, *arXiv:1801.05551*. [Online]. Available: https://arxiv.org/abs/1801.05551

[21] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[22] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

[23] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN—Text conditioned auxiliary classifier generative adversarial network," 2017, *arXiv:1703.06412*. [Online]. Available: https://arxiv.org/abs/1703.06412

[24] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2017, pp. 5907–5915.

[25] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[27] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3668–3678.

[28] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] F. Faghri, D. J. Fleet, J. R. Kiros, G. B. Toronto, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*. [Online]. Available: https://arxiv.org/abs/1707.05612

[31] J. Wang, L. Yu, Y. Xu, Y. Gong, B. Wang, D. Zhang, W. Zhang, and P. Zhang, "IRGAN: A minimax game for unifying generative and discriminative information retrieval models," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 515–524.

[32] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo, "GraphGAN: Graph representation learning with generative adversarial nets," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2508–2515.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2818–2826.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. IEEE Eur. Conf. Comput. Vis.*, May 2014, pp. 740–755.

[36] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3202–3212.

[37] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1346–1353.

[38] R. Jin, J. Y. Chai, and L. Si, "Learn to weight terms in information retrieval using category information," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 353–360.

[39] J. W. Reed, Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson, "TF-ICF: A new term weighting scheme for clustering dynamic data streams," in *Proc. 5th Int. Conf. Mach. Learn. Appl.*, Dec. 2006, pp. 258–263.

[40] S.-K. Song and S. H. Myaeng, "A novel term weighting scheme based on discrimination power obtained from past retrieval results," *Inf. Process. Manage.*, vol. 48, no. 5, pp. 919–930, 2012, doi: 10.1016/j.ipm.2012.03.004.

[41] O. A. S. Ibrahim and D. Landa-Silva, "Term frequency with average term occurrences for textual information retrieval," *Soft Comput.*, vol. 20, no. 8, pp. 3045–3061, 2016.

RINTARO YANAGI received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2019, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interest is in machine learning and its applications.

REN TOGO (S'16–M'19) received the B.S. degree in health sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is currently a Research Fellow of the Japan Society for the Promotion of Science with the Faculty of Information Science and Technology, Hokkaido University. He has a license of Radiological Technologist. His research interest includes machine learning and its applications.

TAKAHIRO OGAWA (S'03–M'08–SM'18) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively, all in electronics and information engineering. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interest is in multimedia signal processing and its applications. He has been an Associate Editor of the *ITE Transactions on Media Technology and Applications*. He is a member of the ACM, EURASIP, IEICE, and the Institute of Image Information and Television Engineers (ITE).

MIKI HASEYAMA (S'88–M'91–SM'06) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively, all in electronics. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor, in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology, Hokkaido University. Her research interest includes image and video processing and its development into semantic analysis. She is a member of the IEICE and Information Processing Society of Japan (IPSJ) and a Fellow of Institute of Image Information and Television Engineers (ITE). She has been the Vice-President of the ITE and the Editor-in-Chief of the *ITE Transactions on Media Technology and Applications* and the Director of the International Coordination and Publicity of the IEICE.